Data Journalism

BRANT HOUSTON
University of Illinois at Urbana–Champaign, USA

Data journalism has been defined as the analysis and visualization of digital information for news stories, news infographics, and interactive presentations. It has used analytic software and methods to detect patterns or outliers within data sets, provide context, or examine trends over time. The digital information may be structured (in columns and rows) or unstructured (text, audio, video).

While the definitions of data journalism have continued to be debated, there has been a developing consensus on a broader definition as suggested by data scientist Alexander Howard. Howard (2014) has defined data journalism as the application of data science to journalism to extract knowledge from data and more specifically as gathering, cleaning, organizing, analyzing, visualizing, and publishing data to support the creation of acts of journalism. As a result of its broad definition, the term data journalism has begun to encompass other terms such as precision journalism, computer-assisted reporting, and computational journalism. An even broader conceptual definition has stated that data journalism is thinking about how to frame an experiment, gather data, and use rigorous methods to build evidence of some finding of journalistic importance.

The different methodologies of analysis in data journalism have included counting and summing by category of information, performing statistical tests to determine likelihood of certain occurrences, creating digital maps, using social network analysis to examine relationships between individuals and institutions, and creating algorithms for tasks such as topic modeling in which possible topics within text are identified by algorithms. The methodologies have also included techniques of gathering and analyzing datasets created from surveys, polling, sensors, crowd sourcing, web scraping, social media aggregation, and field observation. Practitioners and researchers also have suggested that data journalism includes multimedia or interactive presentations of data on the Web or the building software applications for newsrooms for newsgathering or infographics (Berret & Phillips, 2016).

However, some scholars and practitioners have said that data journalism is a part of a sequence of innovation and digital techniques in journalism. The sequence has been identified as developing from precision journalism to computer-assisted reporting to data journalism to computational journalism (Hamilton, 2016). Precision journalism has been defined as the use of social science and statistical analysis for news stories with an emphasis on investigative stories. Computer-assisted reporting has been described as counting, searching, sorting, and cross-indexing, although it has included statistical analysis and visualizing of data. In the past decade some practitioners have argued that data journalism is less focused on traditional journalism, such as investigative and accountability reporting, but more on the content of the data itself and the sharing of

The International Encyclopedia of Journalism Studies. Tim P. Vos and Folker Hanusch (General Editors), Dimitra Dimitrakopoulou, Margaretha Geertsema-Sligh and Annika Sehl (Associate Editors).
© 2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.
DOI: 10.1002/9781118841570.iejs0119

that data with the public and other academic disciplines. In that definition, different kinds of data, such as audio, photos, and videos, have been added to the kinds of data for analysis. Computational journalism has been described as creating ways to employ algorithms for more efficient newsgathering, news writing, and data analysis and presentation, with an emphasis on unstructured big data, machine learning, and emerging technologies to analyze structured and unstructured information to find and tell stories.

There has been general agreement that the practice of data journalism has resulted in news stories and infographics that never could have been done before because of the increased access to large public datasets and the improved expertise by journalists in data science.

It also has been widely accepted that data journalism has transformed the nature of the profession by allowing journalists to gain a deeper and more detailed understanding of the outcomes of laws and policies and to bring context and detail to stories that previously relied on anecdotal cases and isolated occurrences. In addition, data journalism has permitted journalists to share and discuss with the public the data used in new stories and to seek contributions of data from the public (Houston, 2014).

As data journalism has become more routine, it also has given rise to discussion and debate over the issues regarding ethical responsibilities of journalists. Journalists have been criticized for publishing data that violated the privacy of individuals and have been challenged on the ethics of using large datasets leaked by computer hackers or whistleblowers. The use of leaked data has raised questions about the wisdom of publishing of purloined data for news stories, the resulting violation of individual privacy and commercial rights, and potential threats to national security and public safety.

Some practitioners and scholars have suggested data journalism began with the listing and analysis of tabular information in newspapers. They have cited such examples as the publishing of tables of school enrollments and students costs by the British newspaper, *The Manchester Guardian*, in 1821 (Gray, Bounegru, & Chambers, 2012). They also cited an investigation by Horace Greeley's newspaper in the nineteenth century that compared U.S. congressional travel reimbursement rates for mileage to the actual mileage in postal routes to the same locations. The investigation revealed congressmen charging for more mileage than was possible.

But while those examples involved analysis of structured data, other practitioners and scholars have restricted the description of data journalism to practices that involve the use of computers and software to do the analysis and presentation of digital information. As the first example of data journalism, they have identified a CBS news project, which in 1952 used a UNIVAC main frame computer to predict the outcome of United States presidential elections. By analyzing polling data, CBS successfully predicted the result of the election, but did not publish the results until after the election.

Data journalism's actual practical beginning has been viewed as the application of social science methods and statistical analysis by journalist Philip Meyer to examine the causes of the Detroit racial riots in 1967 in the United States. Meyer surveyed city residents who answered questions about participation in riots, attitudes on crime and punishment for looters, and their socioeconomic status. While city officials attributed the riots to migrants from the south of the United States and to the uneducated, Meyer

found a higher percentage of those who grew up in the north to be rioters and also not to the lesser education of rioters (Meyer, 2002).

In 1973, Meyer published the textbook *Precision Journalism: A Reporter's Introduction to Social Science Methods*, which not only provided instruction for doing data analysis, but also called for all journalists to learn data analysis and social science methods. The call for a more analytical approach to reporting echoed the journalist Walter Lippmann, who in the 1920s proposed "objectivity" as a method in journalism that employed social science methods in which a reporter would begin a story with a hypothesis, gather evidence, and present a story based on conclusions from examination of the evidence and not from a predetermined bias.

Meyer continued his data analysis work in the 1970s with *Philadelphia Inquirer* investigative reporters Donald Barlett and James Steele in analyzing sentencing patterns of defendants. During the same decade Rich Morin at *The Miami Herald* analyzed digital property assessment records for news stories.

From the late 1970s to the early 1980s, a small group of U.S. journalists either were taught by Meyer, who offered workshops for professional journalists at the University of North Carolina, or taught themselves how to do data analysis for news stories. In addition to the challenge of learning data journalism while still practicing traditional news gathering, the small group of geographically dispersed reporters had difficulty in getting access to mainframe computers.

A team of investigative journalists in 1976 contracted with a University of Arizona professor to do social network analysis that diagrammed the connections and influence of a powerful group of 40 community leaders in Phoenix, Arizona. The analysis was part of a lengthy investigation known as the Arizona Project into organized crime and government corruption. However, there was not significant use of social network analysis in journalism again until the early twenty-first century.

In the 1980s, Steve Ross, a professor at Columbia University, also offered analytic journalism courses for journalists and journalism students in which he taught computer science and the use of Structured Query Language (SQL). The language did not require sophisticated statistical skills to summarize data or to join datasets together for comparisons, Eliott Jaspin, an investigative journalist from *The Providence Journal-Bulletin* in Providence, Rhode Island, adopted what he learned from Ross during a fellowship at Columbia and began using datasets in his reporting. In 1986, Jaspin published a story that compared a dataset of school bus drivers to a dataset of criminal records and found an alarming number of matches (Jaspin, Johnson, & Fedorzyn-Edgar, 1986).

Jaspin played an additional role in the spread of computer-assisted reporting by easing entry for reporters into the field of data journalism when he and a computer programmer Daniel Woods developed software named Nine Track Express. The software allowed journalists to copy data from nine-track computer tapes on which the data were stored onto a personal computer by using a portable tape drive. The software freed journalists from having to access mainframe computers.

In 1989, the term computer-assisted reporting attained prominence when Jaspin established the Missouri Institute for Computer-Assisted Reporting (MICAR) at the University of Missouri. Although Jaspin taught detailed sessions on the technical

aspects of data, his success stemmed from an approach that relied more on database manager and spreadsheet software rather than statistical software.

In 1989 the story "The Color of Money," won a Pulitzer Prize for the *Atlanta Journal and Constitution* and demonstrated the impact that could be derived from journalists and professors working together. The series of stories relied on analysis of home mortgage data by both an in-house data analyst and university professors. The series of stories exposed racial disparity in mortgage lending and inspired other U.S. newsrooms to begin to adopt data journalism.

During the 1990s, data journalism also continued to strengthen the collaborations between newsrooms and universities as more journalists strived to learn its techniques. In addition to Jaspin's institute, Professor James Brown at Indiana University and Purdue University at Indianapolis (IUPUI) established the National Institute for Advanced Reporting.

The first data journalism conference for journalists was held by Brown in 1990 on the Indiana-Purdue campus and co-hosted by Investigative Reporters and Editors (IRE), a nonprofit association based at the University Missouri. The conference brought together journalists who were practicing or wanting to practice precision journalism or computer-assisted reporting.

After Jaspin left the University of Missouri in 1992, IRE assumed the administration of workshops and the data library that Jaspin created. The change in administration was supported since much of the data journalism being done was by investigative journalists.

In 1994, the institute at Missouri was renamed the National Institute for Computer-Assisted Reporting (NICAR) and IRE raised funds for NICAR staff and activities. NICAR expanded its training beyond national conferences and workshops at Missouri and began conducting workshops throughout the United States and other countries. In 1996, the first widely used textbook *Computer-Assisted Reporting: A Practical Guide* by Brant Houston, who oversaw IRE and NICAR, was published. In 1996 a formal agreement was reached between IRE and the University of Missouri, with the agreement giving IRE formal administrative responsibilities of NICAR. The change from the school's administration to IRE was supported because so much data journalism was being practiced by investigative journalists.

In 1992, the importance of visualizing data was demonstrated by data journalist Steven Doig. Doig, who worked at the *Miami Herald*, showed the efficacy of visualizing data when he collected data and mapped the wind speeds and damage reports after Hurricane Andrew struck Florida in 1992. Doig's analysis and map showed the efficacy of visualization by clearly revealing that buildings received heavy damage from low wind speeds because of poor enforcement of building codes.

As the practice of data journalism increased, journalists began encouraging changes in U.S. open records laws that would recognize that public access to digital documents should be the same as access to paper documents. Government officials had denied access to data, citing privacy and public safety concerns and the difficulty and costs of duplicating the data for journalists. Those arguments were partly overcome when journalists showed that duplication and redaction of records was more accurate and

less costly. In 1996, the U.S. Freedom of Information Act was amended to include digital information. During the same period, states and local governments in the United States also changed their laws to accommodate requests for public data. Laws in other countries later changed as data journalism spread internationally.

With the creation of the World Wide Web in 1989 and the subsequent creation of the Mosaic Web browser in 1993, data became more accessible to journalists as governments and businesses posted datasets on their websites. U.S. journalists began downloading datasets more frequently and the U.S. Bureau of Census was particularly helpful in making data available online in spreadsheets and in geographical information system software for mapping.

The proliferation of data also resulted in stories that were not exclusively investigative and, in a prelude to what became known as data journalism, reporters began producing lifestyle and sports stories based on data to those topics.

In 1996, two Danish journalists, Nils Mulvad and Flemming Svith, attended a NICAR workshop in Missouri and then commenced offering workshops for European journalists at the Danish School of Journalism. They created the Danish International Center for Analytical Reporting (DICAR), wrote a Danish handbook on computer-assisted reporting in 1998, and began organizing conferences on data journalism in Europe. In addition, Mulvad formed a collaborative group of journalists to request and report on farm subsidies in Europe that included an audit of open records laws for digital information in European countries.

In London in 1997, City University journalism professor Milverton Wallace held the first of a series of annual conferences called NetMedia that offered sessions on the Internet and classes in computer-assisted reporting led by NICAR and Danish journalists. The classes covered the basic uses of the Internet, spreadsheets, and database managers, and they were well-attended by journalists from the United Kingdom, other European countries, and Africa. Also in the United Kingdom, the Centre for Investigative Reporting, founded in 1997, began to collaborate with NICAR to offer classes in data journalism during its summer school and has continued a conference with the leadership of both United Kingdom and United States data journalists. By 1999, as a result of the various international training efforts, journalists had produced stories involving data analysis in multiple countries including Finland, Sweden, New Zealand, Venezuela, Argentina, the Netherlands, Norway, Brazil, Mexico, Russia, Bosnia, and Canada.

By the early twenty-first century, data journalism had begun to be practiced by journalists in Africa, Asia, and Australia. The creation of the Global Investigative Journalism Conference in 2001 in Copenhagen included a schedule of lectures and hands-on training sessions in computer-assisted reporting to journalists from nearly 40 countries. Attendance at the biennial conference has increased to include journalists from more than 130 countries. In 2005, Wits University in Johannesburg, South Africa, began an annual investigative conference which provided lectures and hands-on training in computer-assisted reporting to mostly African journalists.

Also in 2005, the term data journalism also began to be used more often by programmers in the United States and in Europe who had become interested in journalism. In the United States, computer programmer and journalist Adrian Holovaty spurred more

interest among programmers when he created from police data an ongoing and interactive map of crime in the city of Chicago that allowed citizens to choose ways to and from work that were supposedly safer. With funding from the John S. and James L. Knight Foundation, Holovaty then created the now-defunct Every Block in 2007, which used local data across the United States for online maps, but the project later was criticized for not checking the accuracy of government data more thoroughly.

Throughout the early twenty-first century, Google online data tools were utilized increasingly by journalists throughout the world because they were free and relatively easy to learn. In 2015, Google created the Google News Lab, now called the Google News Lab – Google News Initiative, to focus on support for journalists using data for news stories. The News Lab also began to offer tutorials for journalists on how to better use Google software and tools.

A free data journalism service for journalists, DocumentCloud, was created with funding from the Knight Foundation in 2009 by Aron Pilhofer at *The New York Times* and Scott Klein and Eric Umansky at the nonprofit newsroom ProPublica and programmer Jeremy Ashkenas. DocumentCloud increased the use of non-structured data by allowing journalists to upload government and business documents in text to the DocumentCloud site. The data was then processed through Thomson-Reuters Open Calais, a software program that converted the text to structured data and returned elements of the text such as entities, topic codes, events, relations, and social tags that made the documents easier to analyze. DocumentCloud also allowed journalists to annotate the documents online.

Despite these collaborative successes, the cultures of journalism and computer programming initially clashed. Programmers believed that journalists were not as transparent with the public as they should be about their analysis of data and that journalists were also reluctant to try new software and approaches. The journalists perceived that the programmers did not respect the need for determining the accuracy and limitations of data sets and did not engage enough in journalism practices, such as interviewing and observations in the field.

By 2009, a desire to bring together the computer programmers and journalists resulted in creation of the group Hacks/Hackers by Pilhofer, Rich Gordon, a professor at Northwestern University, and journalist Burt Herman to bring more sharing and understanding between the two professions and ease some of the culture clash between the groups.

International progress in data journalism continued as the first decade of the twenty-first century was ending. Fred Vallance-Jones and David McKie published *Computer-Assisted Reporting: A Comprehensive Primer* in 2009 with a special emphasis on data in Canada. In 2010, The European Journalism Centre in the Netherlands created another data-driven journalism initiative established at the data journalism website similar to many of the topics and approaches developed by NICAR. A 2011 initiative produced *The Data Journalism Handbook* in which there was a compendium of the points of views and knowledge of practitioners and scholars. The centre also collaborated with the Global Editors Network and Open Knowledge Foundation on data journalism awards.

In 2011, the European Investigative Journalism Center and a group of data journalists organized a conference called Data Harvest to bring together journalists and programmers. Also, in 2011 *The Guardian* in the United Kingdom received much attention for its analysis of the participants in racial riots in cities in the United Kingdom. While it credited part of its analytical approach to Meyer's work, *The Guardian* also did extensive analysis of social media to determine the origins and purposes of the rioters.

The teaching of data journalism became a routine part of journalism conferences in Denmark, Sweden, Norway, Finland, and the Netherlands. Through the global investigative conferences, the use of data quickly spread across Eastern Europe. In Eastern Europe, Drew Sullivan (who formed the Organized Crime and Corruption Reporting Project) and Romanian journalist Paul Radu were strong proponents and practitioners.

In 2010, journalists began to analyze large unstructured data. Wikileaks released the Afghan War Diaries and then the Iraq War Diaries, which contained U.S. secret military documents. It followed those releases with others including state department documents. Other big data leaks have included banking records and records of offshore accounts known as the Panama Papers and the Paradise Papers.

As the access to, and use of, unstructured and big data increased, the need for improved analytic tools and more sophisticated algorithms also grew. In 2008, Georgia Tech hosted a Symposium on Computation and Journalism. Several hundred computer scientists and journalists attended. In 2009, The Center for Advanced Study in the Behavioral Sciences at Stanford University organized a similar workshop and in 2015 Stanford University created the Computational Journalism Lab. Columbia Journalism School began offering a course in computational journalism and the university's Tow Center received a substantial grant from the Knight Foundation for its computational journalism program. In 2015, Syracuse University began a Master's program in computational journalism. The Google News Lab has sponsored awards in computational journalism research and several universities have hosted conferences called the Computation + Journalism Symposium since 2014.

Among the new approaches was one called "algorithmic accountability" in which journalists reverse-engineer algorithms used by social media and online search tools and discover unintended and intended bias and misdirection. There has continued to be a debate on whether all journalists will need to have programming skills but there has been general agreement that all journalists need to know how programmers do their work.

Despite the recognized need for the skills of data journalism, recent surveys have found that many journalists and journalism students still lack the skills to practice it and that universities do not offer enough classes for the field. A survey by educators from Columbia University and Stanford University (Berret & Phillips, 2016) found that half of the journalism schools in the United States did not offer a data journalism class in their curricula. A Google News Lab survey found that about half of newsrooms it surveyed had a dedicated data journalist (Rogers, Schwabish, & Bowers, 2017).

Nonetheless, the practice of data journalism has substantially and continuously increased in scope and depth. The publishing of interactive datasets and visualizations is expected to be much more common. It is predicted that data gathered by drones and sensors will be routinely blended into data journalism as will reporting done with augmented and virtual reality. The use of algorithms to analyze Big Data is expected to increase, as are the collaborations among journalists and university professors.

SEE ALSO: Computational Journalism; Digital Journalism; Hacks and Hackers; Infographics and Data Visualization; Investigative Journalism

References

Berret, C., & Phillips, C. (2016). *Teaching data and computational journalism*. New York, NY: Columbia University.

Gray, J., Bounegru, L., & Chambers, L. (2012). *The data journalism handbook: How journalists can use data to improve the news*. Sebastopol, CA: O'Reilly Media.

Hamilton, J. (2016). *Democracy's detectives: The economics of investigative journalism*. Cambridge, MA: Harvard University Press.

Houston, B. (2014). *Computer-assisted reporting: A practical guide* (4th ed.). New York, NY: Routledge.

Howard, A. (2014). *The art and science of data-driven journalism*. New York, NY: Tow Center for Digital Journalism.

Jaspin, E., Johnson, M. M., & Fedorzyn-Edgar, S. (1986). R.I. System fails to fully check driving records of bus applicants. Investigative Reporters and Editors Inc. Resource Center. Retrieved from https://ire.org/resource-center/stories/6918/

Meyer, P. (2002). *Precision journalism: A reporter's introduction to social science methods* (4th ed.). Lanham, MD: Rowman & Littlefield.

Rogers, S., Schwabish, J., & Bowers, D. (2017). *Data journalism in 2017: The current state and challengers facing the field today*. Mountain View, CA: Google News Lab.

Vallance-Jones, F., & McKie, D. (2009). *Computer-assisted reporting: A comprehensive primer*. Oxford, UK: Oxford University Press.

Further reading

Cairo, A. (2012). *The functional art: An introduction to information graphics and visualization*. Berkeley, CA: New Riders.

Herzog, D. (2015). *Data literacy: A user's guide*. Thousand Oaks, CA: SAGE.

Lathrop, D., & Ruma, L. (2010). *Open government: Collaboration, transparency, and participation in practice*. Sebastopol, CA: O'Reilly Media.

Rogers, S. (2013). *Facts are sacred*. London, UK: Faber & Faber.

Brant Houston is the Knight Chair in Investigative Reporting at the University of Illinois where he works on projects involving the use of data analysis in journalism.

He is co-founder of the Global Investigative Journalism Network and the Institute for Nonprofit News. He is the author of *Computer-Assisted Reporting: A Practical Guide*, co-author of *The Investigative Reporter's Handbook*, and contributor to books on freedom of information and open government. Before joining the University of Illinois, he was the executive director of Investigative Reporters and Editors at the University of Missouri after being an award-winning investigative journalist for 17 years.